Neural Network for Eye Contact Detection

Wall Lab
School of Interactive Computing

Heng Li
Spring 2016

1. Introduction

Eye contact is fundamental to understanding many psychology and cognitive science questions. Human gaze plays an important role in social interaction as it conveys a lot of information in face-to-face communication. Thus, the ability to use computational analysis to identify eye contact in a social interactive environment with high accuracy can facilitate many other research areas.

The problem of understanding eye contact belongs to the bigger problem of understanding gaze. In previous studies, one approach is appearance-based gaze estimation, and most recent works [2][3][5][6][7][8] in this area consist of using a static camera to estimate gaze direction. By obtaining data from a variety of image datasets, the researchers train and develop models to estimate gaze direction [3]. Previous research provides two ways of collecting gaze and head pose pair data for gaze analysis [3][4]. The first is simply obtaining data from a set of calibrated cameras [3]. The second is building a 3D model of a face and generating real time 3D face images as data [4]. So far, the appearance-based gaze estimation approach shows a state-of-the-art performance in gaze estimation.

In my study I proposes an eye contact behavior in a naturalistic social interaction by using a point-of-view (POV) camera [11], which is a 2-3 minutes play interaction between an adult examiner. In this setup, the child is interacting with a social partner, who wears glasses with an outward-facing camera in the bridge over those nose, aligned right between two eyes.

Using video collected in this setup, we apply face detection [12] and facial landmark [12] analysis algorithm to obtain head pose and eye area images. Then we have these two pieces of information as the input to train a convolutional neural network (CNN). In the end, this CNN will serve as a classifier to detect eye contact between the child and the social partner.

2. Methods

2.1 Constructing Dataset

We use the video from the POV camera system [14] as input and detect the child's face in each frame. Then we analyze the facial landmarks of the face to estimate head pose and crop eye area images. Then using these features, we are able to collect a dataset of eye area image – head pose pairs. This dataset contains 23577 pairs of these data.

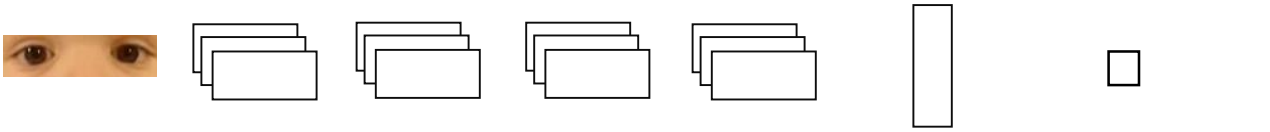
We use the Dlib [12] software to detect the child's face in the egocentric video captured by a camera worn by the child's play partner. For each frame of the video, we run the face detection method by Dlib to estimate a bounding box of the child face. Then we apply face landmark analyzer to the face inside the bounding box to obtain a list of facial landmarks.

Cropped eye area

Using these facial landmarks, we are able to systematically crop the region from the left corner of left eye to the right corner of the right eye, then manually remove those which are incorrectly cropped Then we can also estimate the head pose of the child using the obtained facial landmarks by using Dlib [12]. The head pose consists of three degrees of freedom, which are yaw, pitch and roll.

Each video was annotated by two annotators at the frame level for the onset and offset of each period when the child was making eye contact with the examiner, and each period when the child's face was not visible in the POV video. Our analysis focused only on those frames when the child's face was in view of the POV camera. Thus, we have two kinds of frames: frames when the child's face is visible and the child is making eye contact with the examiner, and frames when the child's face is visible and the child is not making eye contact. We combine this binary eye contact value with the face features of eye area image and head pose. We are able to build a dataset at a size of 23577 frames, with 9736 images of child with eye contact and 13841 images of child with no eye contact.

2.2 Training Convolutional Neural Network



We propose a multimodal CNN modified from the model by Zhang et al [5]. It has max pooling layers between the four convolution layers. In the end there is a fully connected layer. Then we run linear regression together with the head pose to generate the binary classification signal in the end (eye contact vs. no eye contact).

We use Caffe [13] deep learning package to construct this convolutional neural network model, then feed the cropped eye area image data and the head pose direction date to train a classifier for eye contact.

3. Experiments

We collected 10 sessions of the POV child and partner interaction footage. Each of these sessions has a different child and different social situation. Then we pick one of the session as the testing set, and the remaining 9 serve as the training set to train the neural network model.

Our experiment benchmarks the performance of single-frame eye contact detection. This detection is a binary classification problem (eye contact vs. no eye contact) at every frame. We use precision, recall and F1 score as our evaluation criteria.

4. Results

4.1 Face Detection

Session-level results for face detection and eye contact detection can be found in Table 1. The column 'Face Detected' lists the proportion of frames when the child's face is detected by over the total number all frames when the child's face is visible in the POV video. As we can see, the performance of the face detector varies across sessions, as low as 14% and with an overall average detection rate of 55.2%. The failure of face detection directly impacts the step of cropping eye regions and estimating head pose in our pipeline. Thus, when compared with the result from Ye et al [11], our approach has a considerably low recall rate.

Table 1. ADD TITLE

Dataset	Face Detection	Eye Contact Detection: Precision	Eye Contact Detection: Recall	Eye Contact Detection: Accuracy	Eye Contact Detection: F1
1	0.91712	0.98131	0.98348	0.99121	0.98239
2	0.84123	0.98463	0.93759	0.96501	0.96051
3	0.73812	0.98007	0.78405	0.95216	0.87117
4	0.69327	0.98378	0.81523	0.9506	0.89161
5	0.59915	0.97805	0.88135	0.97033	0.92719
6	0.59132	0.9768	0.83159	0.95311	0.89836
7	0.38705	0.98608	0.55048	0.86019	0.70654
8	0.32365	0.99046	0.65995	0.93752	0.79211
9	0.28886	0.98604	0.36208	0.83558	0.52966
10	0.14025	1	0.19721	0.767	0.32805
Average	0.552	0.98472	0.7003	0.91827	0.78876

4.2 Convolutional Neural Network Performance

Table 2. ADD TITLE

	Precision	**Recall**	**F1 Score**
OMRON [14]	0.5151	0.7179	0.5998
Gaze Locking [15]	0.6028	0.6454	0.6234
PEEC [11]	0.7929	**0.7268**	0.7584
Deep Eye Contact	**0.98472**	0.7003	**0.7887**

In Table 2, we report a cross comparison between our approach (Deep Eye Contact) and three previous state-of-the-art approaches. In our analysis, the face detector fails for 44.8% of frames identified as containing eye contact by human annotators (false negative). However, in the previous approach from our group (PEEC [11]), the detector had a much lower missing rate of 17%. Thus, the relatively lower recall rate of our approach is likely due to the face detector.

In terms of precision, our approach significantly out-performs previous state-of-the-art approaches. It also reaches human level performance in terms of precision rate.

5. Discussion

We consider the detection of face and eye areas for eye contact directed from a child to an adult who is wearing a camera. This situation is set up in a natural social interaction scenario. We propose an alternative approach for detecting the events of eye contact in egocentric video based on convolutional neural network.

However, our study on this approach is yet to be concluded. First, we have yet to use a better face detector in our pipeline. Since the failure of detecting face is the main drawback of our pipeline, improving this bottleneck can greatly improve the performance of our approach.

And also, we have collected more videos sessions recently, we believe using more of these sessions to train the neural network model can also help us avoid overfitting. Then after we are able to construct a better pipeline, we can test it on other scenarios to test the performance of our pipeline.

For now, our preliminary experiments show a very promising resulting when using convolutional neural networks to solve this egocentric eye contact detecting problem. We will continue polishing this idea, and hopefully produce a more well-rounded conclusion in the future.

Reference

[1] M. Argyle and J. Dean, “Eye-Contact, Distance and Affiliation.” Sociometry, 1965.

[2] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp. 1148-1161, 1993

[3] Kar-Han Tan, D. J. Kriegman and N. Ahuja, "Appearance-based eye gaze estimation," *Applications of Computer Vision, 2002. (WACV 2002). Proceedings. Sixth IEEE Workshop on*, 2002, pp. 191-195.

[4] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

[5] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, “Appearance- Based Gaze Estimation in the Wild,” in CVPR, 2015

[6] Wood, Erroll, et al. "Rendering of eyes for eye-shape registration and gaze estimation," *arXiv preprint arXiv:1505.05916* (2015)

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 86(11):2278–2324, 1998.

[8] A. Recasens, A. Khosla, C. Vondrick and A. Torralba, “Where are they looking?” *Advances in Neural Information Processing Systems* (NIPS), 2015

[9] Ahmad Humayun, Fuxin Li, and James M. Rehg. RIGOR: Recycling Inference in Graph Cuts for generating Object Regions. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2014.

[10] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn
IntraFace
IEEE International Conference on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia. , 2015

[11] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M. Rehg.
Detecting Bids for Eye Contact Using a Wearable Camera.
11th IEEE International Conference on Automatic Face and Gesture Recognition (FG2015)

[12] Davis E. King. Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research 10, pp. 1755-1758, 2009

[13] Yangqing Jia et al. 2014. Caffe. Proceedings of the ACM International Conference on Multimedia

[14] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In Proceedings of the UbiComp, pages 699–704. ACM, 2012.

[15] B. Smith, Q. Yin, S. Feiner, and S. Nayar. Gaze Locking: Passive Eye Contact Detection for HumanObject Interaction. In ACM Symposium on UIST, pages 271–280, Oct 2013.